The Dynamics and Evolutionary Potential of Domain Loss and Emergence

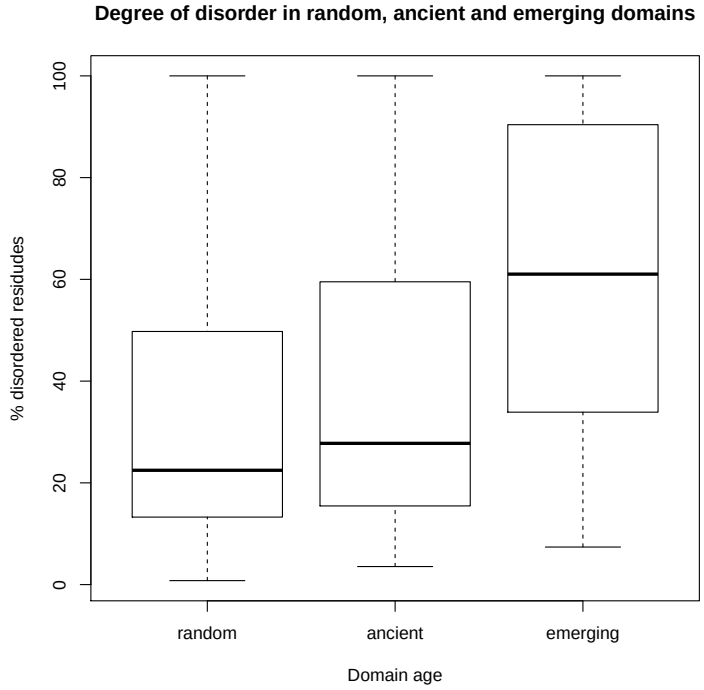
Supplementary Material

Supplementary Table 1: The rate of domain gain and loss per million years along each branch within Arthropods. To obtain the rate along each branch, the absolute number of events (gain and loss) was divided by the respective branch length. Lineage averaged rates were obtained by summing up the absolute values for each event along all branches to the respective species, and dividing by time since LCA.

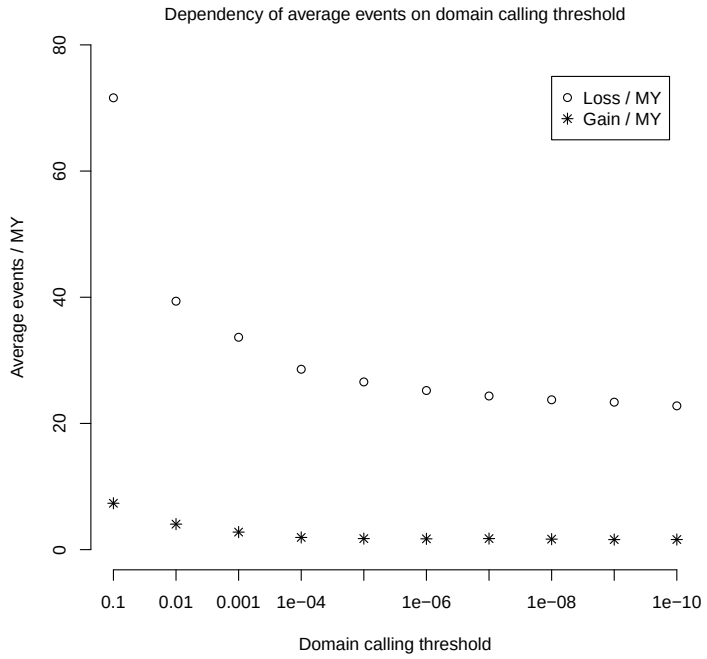
Ancestor node	Descendant node	Gain			Loss		
		Pfam-A	Pfam-B	$\sum$	Pfam-A	Pfam-B	$\sum$
Pancrustacea	D. pulex	0.19	0.28	0.47	0.88	1.90	2.77
Pancrustacea	Endopterygota	0.28	3.52	3.80	0.74	0.62	1.35
Endopterygota	Hymenoptera	0.01	0.06	0.07	1.82	5.12	6.94
Hymenoptera	N. vitripennis	0.21	0.42	0.63	3.05	4.27	7.31
Hymenoptera	A. mellifera	0.16	0.24	0.40	1.67	3.37	5.04
Endopterygota	A	0.90	8.65	9.55	2.50	3.45	5.95
A	T. castaneum	0.06	0.17	0.24	1.04	3.65	4.69
A	B	1.00	7.10	8.10	5.90	6.90	12.80
B	B. mori	0.07	0.20	0.27	1.64	5.17	6.81
B	Diptera	0.31	7.58	7.89	1.33	1.44	2.78
Diptera	Culicidae	0.04	0.24	0.28	1.60	6.97	8.57
Culicidae	Culicinae	0.02	0.10	0.12	0.92	3.40	4.32
Culicinae	A. agypti	0.10	0.35	0.45	1.56	4.02	5.58
Culicinae	C. culex	0.05	0.34	0.39	2.10	2.93	5.03
Culicidae	A. gambiae	0.04	0.27	0.31	1.22	4.27	5.49
Diptera	Drosophila	0.16	2.73	2.89	0.59	0.97	1.56
Drosophila	C	0.00	1.13	1.13	12.63	35.25	47.88
C	D. grimshawi	0.09	0.81	0.91	2.25	8.19	10.44
C	E	0.00	0.50	0.50	2.63	9.13	11.75
E	D. virilis	0.08	1.67	1.75	1.75	7.83	9.58
E	D. mojavensis	0.13	1.83	1.96	1.67	7.46	9.13
Drosophila	Sophophora	0.25	6.00	6.25	5.50	20.75	26.25
Sophophora	D. willistoni	0.14	0.83	0.97	3.86	13.81	17.67
Sophophora	D	0.30	1.20	1.50	1.90	5.40	7.30
D	obscura group	0.08	0.72	0.80	4.40	14.10	18.50
Obscura grp.	D. persimilis	0.84	6.30	7.14	34.45	76.47	110.92
Obscura grp.	D. pseudoobscura	0.84	7.98	8.82	14.29	48.74	63.03
D	Melanogaster grp.	0.17	0.67	0.83	1.08	6.75	7.83
Melanogaster grp.	D. ananassae	0.21	1.50	1.71	8.00	25.43	33.43
Melanogaster grp.	Melanogaster subgrp.	1.75	7.25	9.00	5.25	20.25	25.50
Melanogaster subgrp.	F	0.00	1.50	1.50	34.00	80.50	114.50
F	D. erecta	0.25	1.88	2.13	4.63	14.88	19.50
F	D. yakuba	0.25	2.25	2.50	4.00	15.25	19.25
Melanogaster subgrp.	G	0.00	1.29	1.29	2.14	15.14	17.29
G	D. melanogaster	0.00	2.33	2.33	13.00	58.33	71.33
G	H	1.00	2.00	3.00	61.00	93.00	154.00
H	D. sechellia	0.50	7.00	7.50	30.00	65.50	95.50
H	D. simulans	0.00	5.50	5.50	89.50	194.00	283.50

Supplementary Table 2: GO terms effected by the emergence of novel domains After determining emergent domains within arthropods, all proteins harboring an emerging domain were subjected to a GO analysis using TopGO. By comparison to the set of all arthropodic proteins, over-represented GO terms among the proteins with emergent domains were uncovered. In particular processes related to stimulus response and development frequently acquire emergent domains. Obtained p-values were corrected for multiple testing using Bonferroni, $p < 0.01$ shown.

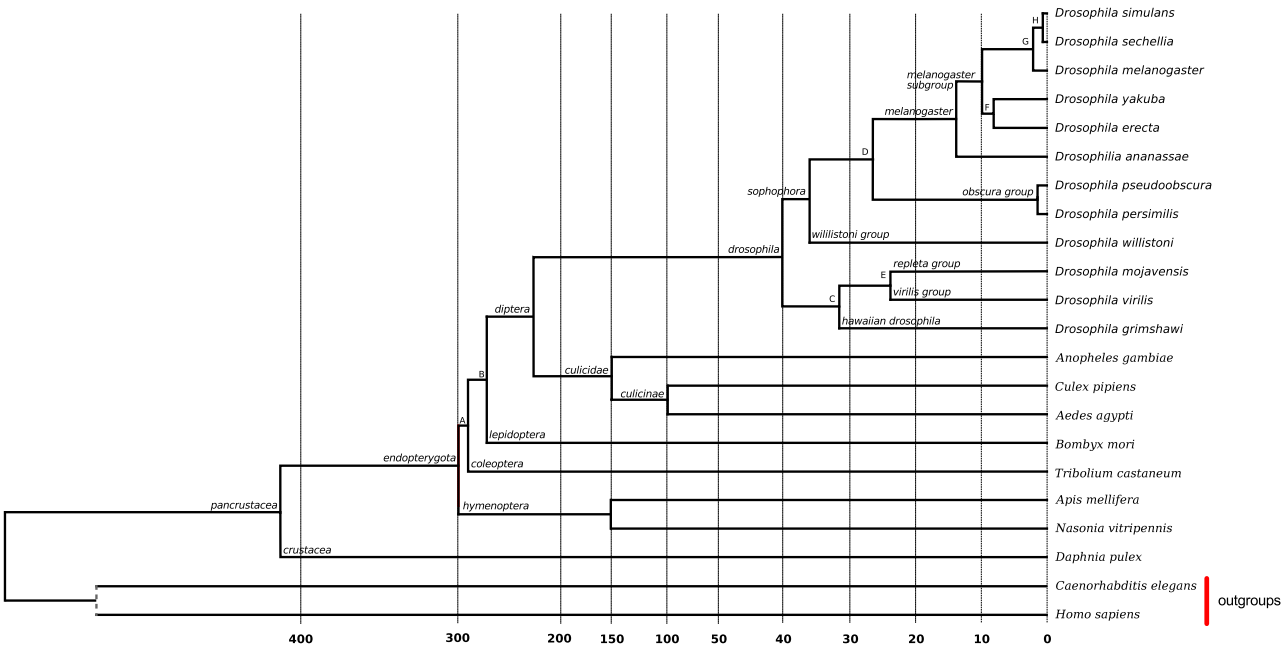
	GO.ID	Term	p
1	GO:0009408	response to heat	4.25E-057
2	GO:0045087	innate immune response	4.95E-057
3	GO:0009411	response to UV	2.39E-041
4	GO:0009414	response to water deprivation	2.57E-037
5	GO:0006979	response to oxidative stress	7.26E-023
6	GO:0009617	response to bacterium	2.40E-022
7	GO:0009409	response to cold	6.95E-019
8	GO:0035110	leg morphogenesis	2.48E-018
9	GO:0042694	muscle cell fate specification	6.10E-018
10	GO:0032275	luteinizing hormone secretion	1.50E-017
11	GO:0007521	muscle cell fate determination	2.64E-017
12	GO:0046884	follicle-stimulating hormone secretion	1.08E-016
13	GO:0035117	embryonic arm morphogenesis	8.60E-016
14	GO:0042048	olfactory behavior	1.29E-014
15	GO:0045662	negative regulation of myoblast differen...	1.75E-014
16	GO:0030540	female genitalia development	3.32E-013
17	GO:0009612	response to mechanical stimulus	8.64E-013
18	GO:0045617	negative regulation of keratinocyte diff...	2.31E-012
19	GO:0008595	determination of anterior/posterior axis...	3.65E-012
20	GO:0030539	male genitalia development	8.05E-012
21	GO:0048738	cardiac muscle tissue development	1.18E-010
22	GO:0042733	embryonic digit morphogenesis	1.97E-010
23	GO:0009314	response to radiation	5.77E-009
24	GO:0030879	mammary gland development	1.69E-007
25	GO:0007569	cell aging	3.91E-007
26	GO:0003007	heart morphogenesis	5.36E-007
27	GO:0045787	positive regulation of cell cycle	7.49E-007
28	GO:0048705	skeletal system morphogenesis	2.58E-006
29	GO:0008544	epidermis development	3.53E-006
30	GO:0045893	positive regulation of transcription, DN...	3.55E-006
31	GO:0051093	negative regulation of developmental pro...	3.76E-006
32	GO:0000122	negative regulation of transcription fro...	7.95E-006
33	GO:0021761	limbic system development	1.02E-005
34	GO:0007350	blastoderm segmentation	1.40E-005
35	GO:0048332	mesoderm morphogenesis	1.85E-005
36	GO:0007026	negative regulation of microtubule depol...	9.09E-005
37	GO:0042770	DNA damage response, signal transduction	1.68E-004
38	GO:0042742	defense response to bacterium	2.56E-004
39	GO:0032880	regulation of protein localization	7.38E-004
40	GO:0009790	embryonic development	1.17E-003
41	GO:0048754	branching morphogenesis of a tube	2.84E-003
42	GO:0016337	cell-cell adhesion	4.93E-003
43	GO:0002376	immune system process	1.10E-002
44	GO:0007617	mating behavior	1.43E-002



Supplementary Figure 1: Degree of disorder in emerging and ancient domains. The 29 domains that emerge within arthropods were grouped into bins according to their age as described in Methods. For each bin, the sequence of all instances of each domain were extracted and the proportion of disordered residues was determined and compared to randomly selected domains. Emerging domains have a significantly higher proportion of disordered residues than ancient domains or random domains (Kruskal-Wallis, $p << 0.001$).



Supplementary Figure 2: Dependency of rates on domain annotation threshold. Domain gain and loss rates vary with different domain E-value thresholds. Loss events exhibits stronger dependency across different E-values; gain rates show comparably little variation as they are restricted by the use of Dollo parsimony. To minimize false-positive domain annotation, we used the model-defined gathering threshold for Pfam-A annotation which varies for each domain. Following previous studies (Ekman, Björklund and Elofsson, 2007), we we used a static cutoff of 0.001 for Pfam-B.



Supplementary Figure 3: Tree of twenty arthropod species with outgroups. To ensure bifurcation, internal nodes were inserted and labeled from A-H, starting with the earliest node. For all calculations, the tree was deeply rooted by inclusion of two outgroups, *H.sapiens* and *C. elegans*. Events along branches to the outgroup were not considered in the analysis. Tree and approximate divergence times based on (HGS Consortium, 2006) and (Hedges, Dudley and Kumar, 2006).